CAN AUDIO ENHANCE VISUAL PERCEPTION AND PERFORMANCE IN A VIRTUAL ENVIRONMENT?

Elizabeth T. Davis, Kevin Scott, Jarrell Pair, Larry F. Hodges, and James Oliverio
Georgia Institute of Technology
Atlanta, Georgia

Does the addition of audio enhance visual perception and performance within a virtual environment? To address this issue we used both a questionnaire and an experimental test of the effect of audio on recall and recognition of visual objects within different rooms of a virtual environment. We tested 60 college-aged students who had normal visual acuity, color vision, and hearing. The between-participants factor was audio condition (none, low fidelity, and high fidelity). The questionnaire results showed that ambient sounds enhanced the sense of presence (or "being there") and the subjective 3D quality of the visual display, but not the subjective dynamic interaction with the display. We also showed that audio can enhance recall and recognition of visual objects and their spatial locations within the virtual environment. These results have implications for the design and use of virtual environments, where audio sometimes can be used to compensate for the quality of the visual display.

Since the early days of Heilig's Sensorama (see Rheingold, 1991) the virtual environment community has been enthusiastic about multi-sensory experiences within a virtual world and the resulting sense of presence or of "being there" (e.g., Dinh et al., 1998). We are concerned with the effect of audio on (a) presence, (b) subjective quality of visual displays, and (c) subjective dynamic interactive realism of the display. We are also interested in whether audio can enhance visually-based memory performance.

Presence can have important implications for real-world problems. For instance, presence can enhance effectiveness of exposure therapy to overcome phobias when the patient is immersed within a simulation of a fearful situation (Rothbaum et al., 1995). Many researchers have found that the addition of audio increases the sense of presence (e.g., Hendrix & Barfield, 1996b; Dinh et al., 1998).

If audio enhances the subjective quality of a visual display, this could be very efficient in the design and use of some multi-modal displays. For instance, in the entertainment industry does incorporating audio into the system compensate for a lower quality visual display? If so, then adding audio is efficient because the information for the high-resolution visual displays requires a much broader bandwidth than that required for audio. A related issue is whether the subjective quality of a visual display is enhanced more by high-fidelity audio (CD quality) than by low-fidelity audio (AM quality). Some have found no effect of audio on perceived realism or quality of the visual display (e.g., Hendrix & Barfield, 1996b) whereas others have found an effect (e.g., Storms, 1998). We evaluated these potential audio effects by using a questionnaire adapted from several other questionnaires reported in the literature (Hendrix & Barfield, 1996a, 1996b; Witmer & Singer, 1998).

In human factors, we are concerned with the relation between perception and performance in designing and testing interfaces. We evaluated whether ambient sounds within a virtual environment could enhance performance on the recall and recognition of visual objects in specific rooms. We also evaluated whether performance was enhanced more by high-fidelity ambient sounds than by low-fidelity sounds. Recall and recognition tests required both correct identification and spatial localization of each object. Although some previous research (e.g., Dinh et al., 1998) suggests that ambient sounds do not enhance memory of a visual object's location, we wanted to more thoroughly test this concept. A perceptually richer environment may help anchor individual objects to specific locations within that environment.

METHOD

Participants

Sixty college-aged students participated in this experiment and received extra credit in their psychology courses for their efforts. Each had a corrected near visual acuity of 20/20 and normal color vision for each eye. All participants reported having normal hearing.

Apparatus and Stimuli

A Pentium Pro 200 MHz computer with a NetPower graphics card, Sound Blaster 16 compatible sound system, and Polhemus InsideTrak tracking system was used to control the visual and auditory displays and to track the participant's head position. The head mounted visual display consisted of an I-glasses personal display system from i-O Display Systems, LLC. This visual display has a horizontal field of view of 30º for each eye and each 0.7" liquid crystal display has a resolution of 180,000 pixels per LCD panel (equivalent to approximately 300 x 200 pixel resolution). The Sony earbud headphones have a frequency response of 16 to 22,000 Hz. A joystick was used to navigate within the virtual environment.

There were four virtual rooms, each with the same spatial layout and decor, but a different wall color (yellow, red, green, or gray). Within each room there was a window, a door, a couch, a painting on the wall, and a bookcase. The bookcase contained outline pictures of seven objects for a total of 28 different objects across the four virtual rooms. These objects were chosen from a standardized set of 260 pictures, with four objects from each of seven categories. The objects within each category were equated for familiarity, image agreement, and name agreement (Snodgrass & Vanderwart, 1980). The seven picture objects within a room came from different categories, so that object category could not facilitate recall of objects seen within the room. In the low- and high-fidelity audio conditions each virtual room also had a distinct ambient sound (i.e., city, ocean, forest, or thunder). The low-fidelity sounds were created with a sampling rate of 11,025 Hz and an 8-bit depth resolution (typical AM quality sound). The high-fidelity sounds were created with a sampling rate of 44,100 Hz and a 16-bit depth resolution (typical CD quality sound). In the no audio condition the earbud headphones served to dampen any occasional sounds within an otherwise quiet testing chamber.

Design

Audio condition (none, low fidelity, and high fidelity) was a between-participants factor. There were 16 combinations of room color (red, yellow, green, and gray) and ambient sounds (city, ocean, forest, and thunder) so that each room color was paired with each sound. Each audio condition used the same 16 combinations. We counterbalanced the order in which rooms were visited across participants for each audio condition.

Procedures

Screening tests. The participants were first tested for near visual acuity using a Tumbling E eye chart, then screened for color vision using Ishihara plates with a C illuminant.

Practice and testing within the virtual environment. The participants were instructed to look around within each virtual room and notice the sights and sounds of that room, including the furniture in the room and the bookcase with pictures of objects. They were told to approach the bookcase, call out the name of each object in the pictures, and informed that later they would be asked questions about the rooms and the things in them. For practice, they were given five minutes to wander within a blue virtual room so that they could learn to navigate through the environment with a joystick. They then were given three minutes in each of the four virtual rooms.

Questionnaire. Afterwards, the participants completed a questionnaire about their experiences within the virtual environment. The questionnaire both provided information and served as a distractor task before the recall and recognition tests. The questionnaire included two questions about a sense of presence, four questions related to static visual 3D realism, two questions about the dynamic realism, and one question about their level of experience with virtual environments. For each answer the participants gave a rating from "1" (e.g., strongly disagree) to "5" (e.g., strongly agree). They had two minutes to complete the questionnaire.

Recall test. Next, each participant recalled the objects seen in each of the four virtual rooms. They also recalled any ambient sounds associated with each room. They had five minutes to complete the recall test.

Forced-choice recognition test. Finally, the participants were given a form with pictures of all 28 objects in one column and the four virtual rooms across the top row. They were instructed to assign each object to the appropriate virtual room. They had five minutes to complete the recognition test.

RESULTS

Questionnaire

For each participant the presence rating was computed by averaging the rating responses for the two presence questions. The static visual 3D and the dynamic realism ratings were similarly computed for each participant by averaging the individual's rating responses for the relevant questions. Orthogonal, planned, focused comparisons were calculated for each of the three ratings. If a significant difference was obtained between the no audio and audio conditions, then a planned comparison was made between the low- and high-fidelity audio conditions.

Participants who experienced ambient sounds within the virtual rooms reported a significantly greater sense of presence (M=3.425, SEM=0.138) than those who experienced only silent visual scenes (M=3.00, SEM=0.192) $t(58)$=1.79, p=0.0345). Moreover, those who experienced ambient sounds also reported a higher degree of static visual 3D realism (M=3.52, SEM=0.066) than those who had no sound (M=3.225, SEM=0.144), although the quality of the visual display was the same for all conditions, $t(58)$=2.12, p=0.038. There was no significant difference in reported dynamic realism between participants who experienced only a visual environment and those who experienced visual rooms filled with ambient sounds. Finally, there was no significant difference between low- and high-fidelity audio conditions for any of these comparisons.

Recall of Objects in Virtual Rooms

Most recalled objects had actually appeared in one of the four virtual rooms. In fact, for all 60 participants a total of only two bogus objects were recalled.

We required participants both to recall each object and to indicate in which of the four virtual rooms it had been seen. For instance, the participant might have recalled seeing a banana in the red room. In this context, a *hit* was a correctly recalled object allocated to the correct room. A *false alarm* was a correctly recalled object misallocated to an inappropriate room or an incorrectly recalled object. From these hits and false alarms we determined a d' sensitivity measure (Macmillan and Creelman, 1991).

Figure 1 shows that the high-fidelity audio condition had more hits, fewer false alarms, and a higher sensitivity measure than either the low-fidelity or no audio conditions. The recall data were noisy, however, so there was only a marginally significant difference between the high-fidelity audio (M=1.825, SEM=0.425) and no audio (M=1.068, SEM=0.407) conditions for the recall d' sensitivity measure, $t(38)$=-1.29, p=0.10.

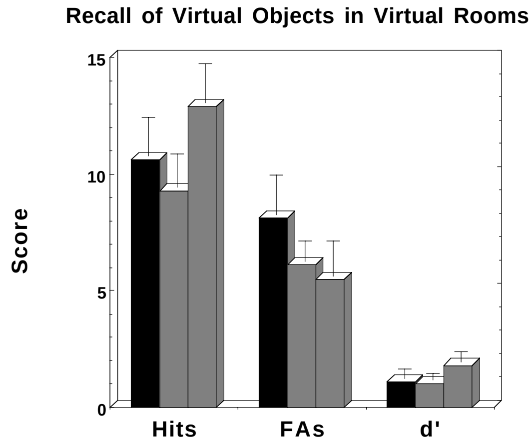

Figure 1
The mean and standard error of the mean are shown for no audio (solid black bars), low-fidelity audio (diagonal stripes) and high-fidelity audio (horizontal stripes).

Forced-choice Recognition of Objects in Virtual Rooms

In the recognition test the participant was shown the objects and asked to indicate in which of four virtual rooms each object had appeared. In this forced-choice recognition test, chance performance corresponded to an accuracy of 25 percent.

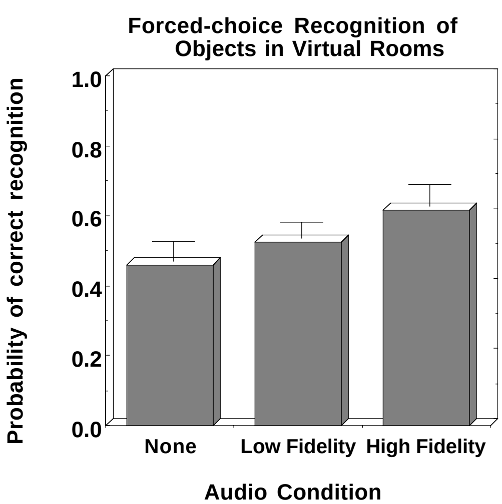

Figure 2
The mean and standard error of the mean for none, low-fidelity, and high-fidelity audio conditions.

The audio condition (none, low fidelity, and high fidelity) had a significant effect on recognition performance, $F(1, 57)=3.95$, $p=0.05$. (See Figure 2.) Forced-choice recognition performance was best for high-fidelity audio ($M=0.6161$, $SEM=0.064$) and worst for no audio ($M=0.4607$, $SEM=0.0546$).

DISCUSSION

Our finding that the addition of audio to the visual display significantly increases the sense of presence is not surprising and agrees with the results of several other multi-modal sensory studies of virtual environments. Nor is it surprising that adding audio to the virtual environment has no effect on the perceived dynamic interaction with the display.

The intriguing result is that the addition of audio can actually enhance the subjective 3D static visual quality of the display, although the physical characteristics of the visual display are the same whether audio is present or not. In a different context, Storms (1998) also reported similar results of audio's effect on visual perception. Our participants reported that the proportions and 3D perspective of visual objects appeared more correct, the depth and volume of the rooms appeared more realistic, the field of view seemed more natural or realistic, and the virtual world appeared more realistic when the visually-perceived rooms were filled with ambient sounds than when they were silent.

These results have implications for design tradeoffs in creating multi-modal VR displays: especially those where the increased sense of "being there" and heightened 3D visual realism matter more than high-quality spatial resolution of the actual visual display. (Note, however, that in some real-world tasks, such as delicate surgical operations, high-resolution visual displays can be critical for good performance.) Because adding audio is computationally less expensive than upgrading the visual quality of the display, the addition of audio can be an efficient way to compensate for low-quality visual characteristics typical of many virtual environments (e.g., those used for entertainment and some forms of psychotherapy).

Our results for recall and forced-choice recognition of objects within virtual rooms also are noteworthy. Why did we find that audio condition has a significant effect on forced-choice recognition of virtual objects in virtual rooms and that high-fidelity audio has a marginally significant effect on recall of objects in the appropriate virtual rooms? Others who have tried (e.g., Dinh et al., 1998) have not necessarily found this relation. First, we tested the recognition and recall of more objects (28 objects) to avoid performance "ceiling effects" and to obtain more reliable measures for each participant. Second, our recall and recognition paradigms are more sensitive than many other paradigms that have been used. We required participants not only to identify the visual objects that they had seen, but also to spatially localize those objects in the appropriate room. In addition, the recognition test was a forced-choice paradigm that has many advantages over yes-no paradigms (e.g., Macmillan and Creelman, 1991). Finally, we used three quality levels of audio (none, low fidelity, and high fidelity) rather than merely the absence or presence of sound. Both recall and forced-choice recognition performance is better with high-fidelity audio than with no audio. If we had used only low-fidelity audio versus no audio, however, we would have concluded that audio has no significant effect on recall or recognition – similar to the conclusions drawn by Dinh et al. (1998). In any case, our results suggest that enriching the perceptual environment by combining audio and visual information can enhance memory performance which requires both identification and spatial localization of visual objects.

ACKNOWLEDGEMENT

We wish to thank Ron Raymond for technical assistance in creating the ambient sounds. We also appreciate the time and effort of all the participants in this study.

REFERENCES

Dinh, H.Q., Walker, N., Song, C., Kobayashi, A., & Hodges, L.F. (1998). Evaluating the importance of multi-sensory input on memory and the sense of presence in virtual environments. *IEEE Proceedings of Virtual Reality 99.*

Hendrix, C. & Barfield, W. (1996a). Presence within virtual environments as a function of visual display parameters. *PRESENCE, 5*(3), 274-289.

Hendrix, C. & Barfield, W. (1996b). The sense of presence within auditory virtual environments. *PRESENCE, 5*(3), 290-301.

Macmillan, N.A. & Creelman, C.D. (1991). *Detection Theory: A User's Guide.* New York: Cambridge University Press.

Rheingold, H. (1991). *Virtual Reality.* New York: Summit Books.

Rothbaum, B.O., Hodges, L.F., Kooper, I.R., Opdyke, M.S., Williford, J.S., & North, M.S. (1995). Effectiveness of computer generated (virtual reality) graded exposure in the treatment of acrophobia. *American Journal of Psychiatry, 152,* 626-628.

Snodgrass, J.G. & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 174-215.

Storms, R.L. (1998). Auditory-visual cross-modal perception phenomena. (Doctoral dissertation, Naval Postgraduate School, 1998.) *Dissertation Abstracts International.*

Witner, B.G. & Singer, M.J. (1998). Measuring presence in virtual environments: A presence questionnaire. *PRESENCE, 7,* 225-240.